ESANN'2009 proceedings, European Symposium on Artificial Neural Networks - Advances in Computational Intelligence and Learning. Bruges (Belgium), 22-24 April 2009, d-side publi., ISBN 2-930307-09-9.

Classification of high-dimensional data for cervical cancer detection

Charles Bouveyron[1], Camille Brunet[2*] and Vincent Vigneron[2]

1- SAMOS-MATISSE, CES, UMR CNRS 8174 – University Paris 1
90 rue de Tolbiac – 75013 PARIS - FRANCE

2- IBISC TADIB, FRE CNRS 3190 - University of Evry
40 rue Pelvoux CE 1455 – 91020 EVRY – FRANCE

Abstract. In this paper, the performance of different generative methods for the classification of cervical nuclei are compared in order to detect cancer of cervix. These methods include classical Bayesian approaches, such as Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) or Mixture Discriminant Analysis (MDA) and a high-dimensional approach (HDDA) recently developed. The classification of cervical nuclei presents 2 main statistical issues, scarce population and high-dimensional data, which impact on the ability to successfully discriminate the different classes. This paper presents an approach to face the problems of unbalanced data and high-dimensions in the context of cervical cancer detection.

1 Introduction

In Statistics, the success of a classification is based on two objectives: phenomenon explanation and phenomenon prediction. In supervised classification, lots of works have been already done and applied with success on different fields. In particular, generative approaches, in which each class is modeled by a known or unknown density function, have been tested in many application fields with success. This study aims to test different methods of generative approach on a public health field: cervical cancer detection. This work emphasizes the discrepancies between theory and the real world, and allows to find new challenges for Statistics in this particular applied field. In this paper, traditional generative approaches and a recently proposed method designed for high-dimensional data [1] will be confronted with the task of cervix cancer cell detection. The detection of abnormal nuclei stands for three important statistical problems. The first one is because of the *scarcity* of the studied population (abnormal cells), the second problem is linked to the *heterogeneity* of the population and the last problem concerns the *high dimensionality* of the data since each nucleus is described by more than 100 features. In the literature, many methods have been proposed to deal with this last problem such as [2]. To face this problem, two different approaches are considered here: feature selection combined with Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) or Mixture Discriminant Analysis (MDA) and the use of High Dimensional Discriminant Analysis (HDDA) [1] which uses all dimensions for classification.

*corresponding author: camille.brunet@ibisc.univ-evry.fr.

ESANN'2009 proceedings, European Symposium on Artificial Neural Networks - Advances in Computational Intelligence and Learning. Bruges (Belgium), 22-24 April 2009, d-side publi., ISBN 2-930307-09-9.

This paper is organized as follows. In Section 2, the dataset and the cytological background are described. Section 3 presents the generative methods used for cells classification. The results of the study are presented in Section 4 and the last section outlines open problems and questions.

2 Cytological background and dataset

The dataset comes from a study on the performance of the DNA ploidy measurements on cervical nuclei for the detection of abnormal cells [3]. For the study, 20 samples of smears containing between 2,000 and 4,000 cells are used. Each nucleus is described by 111 features which may be divided into 3 categories: morphological, photometric or texture features. These nuclei can be divided in 3 classes: normal nuclei, uncertain nuclei which are infected but not diseased and the diseased nuclei. The diseased population is really scarce in our dataset and represents only 0.5% of the population studied compared to the 92.9% of normal nuclei and 6.6% of uncertain nuclei. The diagnosis of each cell has been provided by a gynecological pathologist.

In this study, since we want to keep an individual analysis, the data are split into a learning and a test sets according to a jackknife resampling approach: 19 individuals are used for learning and the 20th is used for the test. The performance of each classifier is evaluated by two medical terms on the testing set: false negative (FN) rate and false positive (FP) rate. The FN rate is defined as the ratio between false-negatives (abnormal cells misclassified) divided by the total number of nuclei with disease. The FP rate is defined as the false-positives (normal nuclei and uncertain nuclei misclassified) over the total nondiseased. The Receiver Operating Characteristic (ROC) curve, which depicts a trade-off between True Positive rate and FP rate, is commonly used in medicine. But since the goal of this study is to detect all diseased nuclei, the main criterion chosen to evaluate the performance of a classifier is the FN ratio. The results presented in Section 4 are the average of FP and FN ratio on the 20 test samples. Thus, each individual has the same weight in the average of FN or FP rate regardless of the number of cancer nuclei.

3 Generative approach for cervical nuclei classification

This study aims to compare two different approaches of the classification: conventional classification methods on selected features and a high-dimensional classification method (HDDA). Both approaches are briefly reviewed below.

3.1 Usual classification methods on selected features

Feature selection In the considered application, a large number of features characterize a nucleus. It is well-known that traditional Bayesian classifiers suffer from the *curse of dimensionality* mainly due to high-dimensional data. Therefore a preliminary step of dimension reduction is required. A comprehensive overview of many existing methods of feature selection has been written by Dash and

Liu [4]. In this work, we use the Wilks' lambda measure which is defined by the ratio between the determinant of the within-covariance and the determinant of the total covariance. This measure is combined with a forward procedure. This procedure begins with an empty set of variables: at each step the variable which contributes the most to the discriminatory power of the model (measured by Wilks' lambda) is included. The selection process stops when no more variable improve the Wilks' lambda.

Generative classification methods commonly used Statistical discriminant analysis methods such as LDA, QDA and MDA arise in a Gaussian mixture model. These methods are concerned with the construction of a statistical decision rule which allows to identify the population membership of an observation. The predicted class is chosen to maximize the posterior class probability given the observation. The main assumption of these methods concerns the distribution of the observations of each class which is Gaussian. We refer to Chapter 4 of [5] for details on LDA and QDA. MDA, developed by Hastie and Tibshirani [6], is a generalization of LDA in which each class is modeled by a mixture of Gaussians. This modeling gives more flexibility in the classification rule than LDA and allows MDA to consider heterogeneity in a class.

3.2 High Dimensional Discriminant Analysis

High Dimensional Discriminant Analysis (HDDA) has been proposed by Bouveyron *et al.* [1] and links regularization, parsimonious models and the idea of dimension reduction. This recent discriminant analysis method is based on the idea that high-dimensional data live in low dimensional subspaces and that data from different classes could live in different subspaces with different intrinsic dimensions. The background is the same as the previous generative approaches. HDDA assumes that each class is modeled by a Gaussian density function and puts constraints on the covariance matrices by taking account of the intrinsic dimension of each class. In addition, it is possible to make additional assumptions on the model to further limit the number of parameters to estimate. In the context of cervical cells classification, HDDA is attractive for two main reasons. First, HDDA does not require a feature selection step and directly models the data in a low dimensional space. The advantage of such an approach is that no information is lost. Second, HDDA could provide posterior information about feature extraction and this could help the practician to understand the resulting classification.

4 Experimental results

This study aims to compare four generative methods of classification according to two different approaches. The first approach aims to directly discriminate the 3 classes whereas the second one is made of two classification steps.

methods	FN	FP
FS+LDA	0.1998 *(0.2157)*	0.00001 *(0.0008)*
FS+QDA	0.1859 *(0.2988)*	0.00013 *(0.0001)*
FS+MDA	0.3267 *(0.3968)*	0.00001 *(0.0009)*
HDDA	**0.1421** *(0.0972)*	0.0084 *(0.0001)*

Table 1: Means and standard deviations (in parentheses) of FN and FP rates for 3-class classification (normal, uncertain and diseased nuclei).

4.1 A 3-class classification

This first approach of nucleus classification is intuitive since a new observation is assigned to one of the normal, uncertain or diseased classes in one forward direct step. As it can be seen in Table 1, HDDA seems to be relatively performant: 14.21% of diseased cells are misclassified which is a good enough performance. However, diseased cells are not well detected for a classical generative approaches and the results are very unstable. The main reason of such a misclassification of abnormal cells can be explained by unbalanced groups. Indeed, nuclei groups studied are not equally represented: only 0.25% of nuclei is diseased. The problem of cancer cells scarcity mainly operates on the feature selection since the selection of relevant variables subset aims to discriminate the different classes. However, because of the low proportion relatively to other classes, cancer nuclei are not representative, as such the results are not satisfactory. To face this scarcity, minority class need to be oversampled and the majority class need to be undersampled as proposed in [7] in order to obtain better classification results. Therefore, the next section describes a two-step classification approach which allows, in a certain way, to "oversample" the minority class.

4.2 A two-step classification

In the aim of resolving the problem of unbalanced data, a two-step classification is now proposed: the first step classifies normal and abnormal [1] nuclei and the second step screens only abnormal elements in uncertain or diseased cell class.

abnormal nuclei combine uncertain and diseased nuclei.

First step (normal against abnormal nuclei) This first step looks for screening the observations in normal or abnormal cell class and is preceded by feature selection (FS) for LDA, QDA and MDA. Among the total number of features, 5 variables are selected according to the Wilks' lambda decreasing. The main variables selected are texture and photometric characteristics. Regardless of the performance criteria, this one-step classification is really performant for each method (see Table 2). The FN criterion arises up to 0.00039 and the FP rate is lower than 0.0710 which is very satisfying compared to the state-of-the-art results. HDDA seems to be relatively less performant than the other methods

[1] Abnormal nuclei combine uncertain and diseased nuclei

methods	FN	FP
FS+LDA	0.0037 *(0.0103)*	0.0006 *(0.0013)*
FS+QDA	$< \mathbf{1.10^{-4}}$ *(0.0061)*	0.0040 *(0.0061)*
FS+MDA	0.0034 *(0.0101)*	0.0007 *(0.0014)*
HDDA	0.0039 *(0.0021)*	0.0705 *(0.0320)*

Table 2: Means and standard deviations (in parentheses) of FN and FP rates for the 1st step of normal and abnormal nuclei classification.

methods	FN	FP
FS+LDA	0.2470 *(0.3281)*	0.0001 *(0.0000)*
FS+QDA	**0.0350** *(0.0631)*	0.0072 *(0.0320)*
FS+MDA	0.2106 *(0.3006)*	0.0003 *(0.0010)*
HDDA	0.1421 *(0.0972)*	0.0801 *(0.0145)*

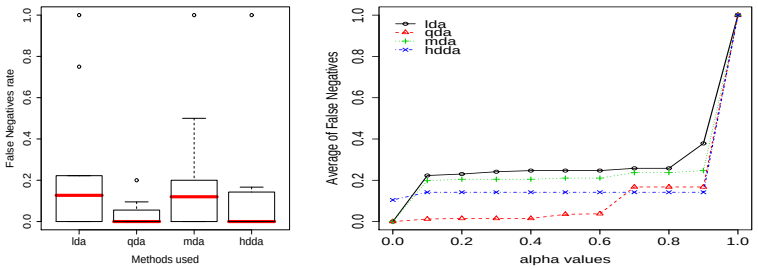
Table 3: Means and standard deviations (in parentheses) of FN and FP rates for the 1st step and the 2nd step of diseased and uncertain nuclei classification.

on the FP criterion and the best method in terms of sensitivity is QDA which detects all abnormal nuclei.

Second step (uncertain against diseased nuclei) In this second step, only diseased and uncertain cells are considered. As previously, before using LDA, QDA or MDA, a preliminary feature selection (FS) was executed and 8 texture variables were selected according to the Wilks' Lambda. First, QDA produces a low and stable FN rate compared to MDA or LDA (see Table 3) and detects all cancer nuclei for 50% of individuals (Figure 1.a). Second, HDDA provides a relative low FN rate as well, even though resampling brings no improvement of FN rate. However, this results improvement for QDA and MDA can be explained by feature selection which performs well because of a better representation of cancer nuclei class in learning sets. Moreover, figure 1.b presents the average FN rate against the classification threshold variation for the studied methods. MDA and LDA still remain the less performant approaches whatever the threshold value is. As previously, the best method for diseased nuclei detection is QDA even though the decision threshold changes (figure 1.b). Finally, after the two classification steps, QDA misclassifies only 1.64% of cancer cells.

5 Conclusion and discussion

To summarize, this study has highlighted some limitations of usual classification approach in the context of cervical cancer detection and proposed an approach to overcome them. First, it has shown that the selection of relevance variables subsets depends on the class size. In classical approaches (LDA, QDA and MDA), the scarcity impacts on feature selection and therefore on classification performance. In such a case, the criterion of feature selection is not able to



(a) (b)

Fig. 1: (a) Boxplots of 10 tests sets FN rates (b) Curves of FN rates evolution according to α-values (right).

discriminate a large class from a small one. Therefore, this paper has proposed a method for feature selection followed by classification which misclassifies less than 4 nuclei on 100. This 2-step approach deals with the scarcity problem and enables to oversample the minority class improves the misclassification rate for QDA and MDA. Second, the study has demonstrated that a high-dimensional approach (HDDA) which has no need for preliminary feature selection can be successfully used in such a context and is robust whatever is the classes proportion of the studied population. Finally, HDDA and QDA work well for cancer cells detection. From these preliminary results, further works could consist in vector quantization of the majority class, so that the final number of code vectors (almost) would equalize the size of the minority class. Then feature selection could be applied on these code vectors and the other class. We must outline that this study has focused only on the nuclei classification. An outstanding challenge remains: how to handle automatically with heterogeneous junks which represent more than 40% of elements contained in a smear image?

References

[1] C. Bouveyron, S. Girard, and C. Schmid. High dimensional discriminant analysis. *Communication in Statistics: Theory and Methods*, 52(1):502–519, 2007.

[2] J.H. Friedman. Regularized discriminant analysis. *The journal of the American statistical association*, 84:165–175, 1989.

[3] M. Guillaud, J.L. Benedet, S.B. Cantor, G. Staerkel, M. Follen, and C. MacAulay. Dna ploidy compared with human papilloma virustesting (hybrid capture II) and conventional cervical cytology as a primary screening test for cervical high-grade lesions and cancer in 1555 patients with biopsy confirmation. *Cancer*, 107(2), 2006.

[4] M. Dash and H. Liu. Feature selection for classification. *Intelligent Data Analysis*, 1:131–156, 1997.

[5] C. Bishop. *Pattern recognition and machine learning*. Springer, New York, 2006.

[6] T. Hastie and R. Tibshirani. Discriminant analysis by gaussian mixture. *Journal of the Royal Statistical Society*, 58(1):155–176, 1996.

[7] N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of artificial Intelligence Research*, 16:321–357, 2002.